

# 模糊半监督加权聚类算法的有效性评价研究

李龙龙<sup>1,2,3</sup>, 何东健<sup>2</sup>, 王美丽<sup>4</sup>

(1. 陕西工业职业技术学院 信息工程学院, 陕西 咸阳 712000;

2. 西北农林科技大学 机械与电子工程学院, 陕西 杨凌 712100;

3. 英国诺丁汉大学 计算机学院, 英国 诺丁汉郡 NG81BB;

4. 西北农林科技大学 信息工程学院, 陕西 杨凌 712100)

**摘要:** 鉴于最佳聚类数在提高聚类算法性能并扩大其应用领域方面的重要性, 为了有效解决聚类算法中最佳聚类数的确定问题, 解决传统的聚类分析算法常常需要人为预先指定聚类数的缺点, 文中提出一种新型模糊半监督加权聚类算法。首先使用该算法对实测数据进行聚类, 获取聚类结果。随后采用4种模糊聚类有效性评价算法依次对不同聚类数下的聚类结果进行聚类分析, 最终通过不同聚类评价结果的对比分析得到实验数据的最佳聚类数。自测数据集的相关实验结果表明, 不同的聚类有效性评价算法具有不同的优缺点, 选择合适的聚类评价算法能够有效地解决最佳聚类数的确定问题, 并能够有效提高实测数据的聚类识别率。

**关键词:** 聚类有效性; 半监督聚类; 算法评估; 成对约束; 最佳聚类数

中图分类号: TP182

文献标识码: A

文章编号: 1673-629X(2016)06-0065-04

doi: 10.3969/j.issn.1673-629X.2016.06.014

## Study of Clustering Validity Evaluation on Semi-supervised Clustering Algorithm with Feature Discrimination

LI Long-long<sup>1,2,3</sup>, HE Dong-jian<sup>2</sup>, WANG Mei-li<sup>4</sup>

(1. College of Information Engineering, Shaanxi Polytechnic Institute, Xianyan 712000, China;

2. College of Mechanical & Electronic Engineering, Northwest A & F University, Yangling 712100, China;

3. School of Computer Science, University of Nottingham, Nottingham NG81BB, UK;

4. College of Information Engineering, Northwest A & F University, Yangling 712100, China)

**Abstract:** As the optimal clustering number has great importance in improving the performance of clustering algorithm and expanding the algorithm's application area, in order to solve the problem of the determination of the optimal clustering number for clustering algorithms effectively and settle the problem that the traditional clustering algorithm often requires prespecified number of clustering, a novel semi-supervised fuzzy clustering algorithm with feature discrimination (SFFD) is proposed. Firstly, it is used to obtain the clustering result of the measured data, and then four kinds of fuzzy clustering validity evaluation algorithm are adopted for clustering analysis under different clustering number. Finally, by the comparative analysis of various validity evaluation algorithm with experimental data the optimal clustering number was obtained. The experiment based on self-test datasets shows that various clustering validity evaluation algorithm has both the advantages and disadvantages, making a good choice for the clustering validity evaluation algorithm can effectively handle the problem of the determination of the optimal clustering number and enhance the recognition rate effectively for the measured data.

**Key words:** clustering validity; semi-supervised clustering; algorithm evaluation; pairwise constraints; optimal clustering number

## 0 引言

作为一种机器学习、数据挖掘领域中常见的数据

分析手段和工具<sup>[1]</sup>, 聚类分析的目标是寻找并发现隐含在输入数据集中具有相似特征的数据集, 即称为簇

收稿日期: 2015-08-07

修回日期: 2015-11-11

网络出版时间: 2016-05-05

基金项目: 国家“863”高技术发展计划项目(2013AA10230402); 国家自然科学基金资助项目(61402374); 陕西工院科研项目(ZK11-34)

作者简介: 李龙龙(1983-), 男, 讲师, 博士, 英国访问学者, 研究方向为智能化检测与技术、智能信息系统; 何东健, 教授, 博士生导师, 研究方向为智能化检测与控制、农业信息技术等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160505.0828.066.html>

的元素集合<sup>[2]</sup>。而聚类问题由于没有事先定义的分类模型或实例来表明不同元素的何种聚类结果是符合预期的,加之分类结果的不可预知性,使得传统聚类算法的评价多来自猜测和假设<sup>[3]</sup>。如何对一个聚类结果及其有效性进行较为全面客观的评判,是一个既复杂又十分困难的技术难题。

常见的聚类评价算法有内部评价法、外部评价法、相对评价法<sup>[4-5]</sup>及模糊聚类有效性评价法<sup>[4-6]</sup>等。其中,内部和外部评价法都基于计算复杂度较高的统计测试,其有效性指标是用来衡量输入数据集与事先已知结构的匹配程度。相对评价法则旨在探索某一聚类算法在特定的假设及参数下能够获得的最佳聚类结果。对于模糊聚类算法而言,模糊聚类有效性评价法则是其最有效的评价算法。而在现有聚类评价算法中,有些聚类有效性评价指数能够求出最佳聚类数<sup>[6-9]</sup>,从而有效解决聚类预设参数中聚类数的确定问题。

考虑到不同聚类评价算法的适用范围,文中给出一种特征加权的模糊半监督聚类算法(SFFD)<sup>[10]</sup>。该算法基于完全自适应距离函数、特征加权<sup>[11-12]</sup>和成对约束构建统一目标函数,用来搜索成对约束下的最优原型参数及最优特征权集。同时,给出四种模糊聚类有效性评价算法,通过不同算法对 SFFD 算法进行有效性评价,进而得出不同输入数据集的最佳聚类数,从而确定聚类过程中的聚类数。

## 1 特征加权的模糊半监督聚类算法

SFFD 算法旨在搜索成对约束下的最优模型参数和最优特征权重集合,其主要算法的公式如下所述。

(1) 聚类之间的距离公式:采用内积范式  $A_i$  来检测数据集中不同聚类的几何形状。

$$d_{ijk}^2 = (x_k - c_i)^T A_i (x_k - c_i) \quad (1)$$

$$c_i = \frac{\sum_{j=1}^N (u_{ij})^m x_{ij}}{\sum_{j=1}^N (u_{ij})^m} \quad (2)$$

$$A_i = (\rho_i \det(F_i))^{-\frac{1}{m}} F_i^{-1} \quad (3)$$

$$F_i = \frac{\sum_{j=1}^N (u_{ij})^m (x_j - c_i)(x_j - c_i)^T}{\sum_{j=1}^N (u_{ij})^m} \quad (4)$$

式中,  $c_i$  为聚类均值,是实例  $i$  对于聚类  $j$  的隶属度。

(2) 特征权值  $v_{ik}$  可以表示如下:

$$v_{ik} = \frac{1}{n} + \frac{1}{2\delta_i} \sum_{j=1}^N (u_{ij})^2 \left[ \frac{\|x_j - c_i\|^2}{n} - d_{ijk}^2 \right] \quad (5)$$

$$\delta_i^{(t)} = K \frac{\sum_{j=1}^N (u_{ij}^{(t-1)})^2 \sum_{k=1}^n v_{ik}^{(t-1)} (d_{ijk}^{(t-1)})^2}{\sum_{k=1}^n (v_{ik}^{(t-1)})^2} \quad (6)$$

其中:  $n$  为输入数据集的特征数;  $K$  为一个常量; 带有上标  $(t-1)$  的变量  $u_{ij}, v_{ik}, d_{ijk}$  分别对应其在第  $(t-1)$  次迭代中的值。

(3) 引入成对约束并采用拉格朗日乘数法进行推导,可以得到算法的目标函数:

$$J_2 = J_1 + \alpha \left( \sum_{(x_i, x_j) \in M} \sum_{p=1}^C \sum_{l=1, l \neq k}^C u_{ip} u_{jl} + \sum_{(x_i, x_j) \in \zeta} \sum_{p=1}^C u_{ip} u_{jp} \right) - \varepsilon_I \left( \sum_{k=1}^C u_{ik} - 1 \right) \quad (7)$$

$$J_1 = J + \sum_{i=1}^C \delta_i \sum_{k=1}^n v_{ik}^2 - \sum_{i=1}^N \lambda_i \left( \sum_{k=1}^n v_{ik} - 1 \right) \quad (8)$$

$$J = \sum_{i=1}^C \sum_{j=1}^N u_{ij}^m \left( \sum_{k=1}^n v_{ik} d_{ijk}^2 \right) \quad (9)$$

(4) SFFD 算法的实例隶属度值可以表示为:

$$u_{rs} = \frac{\varepsilon_I}{2v_{rk} d_{rsk}^2} - \frac{\alpha \left( \sum_{(x_i, x_j) \in M} \sum_{l=1, l \neq s}^C u_{jl} + \sum_{(x_i, x_j) \in \zeta} u_{js} \right)}{2v_{rk} d_{rsk}^2} \quad (10)$$

$$\varepsilon_I = \frac{2}{\sum_{k=1}^C \frac{1}{v_{rk} d_{rsk}^2}} + \alpha \frac{\left( \sum_{(x_i, x_j) \in M} \sum_{l=1, l \neq s}^C u_{jl} + \sum_{(x_i, x_j) \in \zeta} u_{js} \right)}{\sum_{k=1}^C \frac{1}{v_{rk} d_{rsk}^2}} \quad (11)$$

其中:  $M$  为 must-link 约束集;  $\zeta$  为 cannot-link 约束集。

## 2 模糊聚类评价算法

为了更为准确地获取输入数据集的聚类数,可以人为设定不同的聚类数并采用不同的聚类有效性算法对获得的模糊分割矩阵的优劣进行评估,进而得到最佳聚类数。由于现有评价算法各自有不同的缺陷,单一的评价算法无法获得较为可靠的结果,因此,给出了四种不同的聚类结果评价算法来进行综合评价:

(1) 分配系数(PC):由 Bezdek 等<sup>[13]</sup>给出定义,用来测量不同聚类之间的重叠程度:

$$PC(c) = \frac{1}{N} \sum_{i=1}^c \sum_{j=1}^N (\mu_{ij})^2 \quad (12)$$

式中:  $N$  为输入数据集中的实例数目;  $c$  为聚类数;  $\mu_{ij}$  为数据点  $j$  对于聚类  $i$  的隶属度。

当聚类数为最佳聚类数的时候,该系数为其所有取值的最大值。该系数的缺陷是其取值会随着聚类数  $c$  的减少而单调递减,并且其与输入数据集结构之间的关系较为松散。

(2) 分类熵(CE):该系数与 PC 类似,其常用来测

量聚类分割的模糊性:

$$CE(c) = \frac{1}{N} \sum_{i=1}^c \sum_{j=1}^N \mu_{ij} \log(\mu_{ij})$$

(13)

该系数取值会随着聚类数  $c$  的增加而单调递增,并且其与初始输入数据集的关系不是很密切。

(3)分割指数(SC):是指聚类紧密度之和与其间距的比率。该系数是一种基于模糊基数(模糊集的势)的单簇聚类有效性之和<sup>[14]</sup>:

$$SC(c) = \frac{\sum_{j=1}^N (\mu_{ij})^m \|x_j - v_i\|^2}{N_i \sum_{k=1}^c \|v_k - v_i\|^2}$$

(14)

当聚类数为最佳值时,该系数取其最小值。

(4)谢和贝尼指数(XB):该系数可表示为聚类内全变差与聚类间距的比率<sup>[15]</sup>,公式如下:

$$XB(c) = \frac{\sum_{i=1}^c \sum_{j=1}^N (\mu_{ij})^m \|x_j - v_i\|^2}{N \min_{i,j} \|x_j - v_i\|^2}$$

(15)

当其取值为最小值时,聚类数为最佳。

3 实验结果

3.1 数据介绍

为了分析不同的模糊聚类有效性评价算法在确定输入数据集最佳聚类数上的优缺点,并检测文中算法在实际应用中的效果,采集了10种树木在不同时期的160张叶片的照片,每张照片获取其Margin、Shape、Texture及Combination特征作为不同的输入数据集,这些数据集集中的数据均以数值形式存在,其结构如表1所示。

表1 文中采用的数据集

数据集名称	属性数	类数	样本数
Margin	16	10	160
Shape	32	10	160
Texture	8	10	160
Combination	64	10	160

3.2 最佳聚类数的确定

通常大多数聚类算法需要用户预先输入希望产生的聚类数,这就会人为地产生误差且使得结果具有一定的主观性。为了测试确定不同输入数据集的最佳聚类数,分别使用Margin、Shape、Texture及Combination等测试数据作为输入数据集,聚类数  $c$  的预设范围为2~20,采用指数PC、CE、SC和XB对其SFFD聚类结果进行有效性评价分析,结果如图1所示。

图1为不同特征输入数据集在SFFD聚类算法下4种聚类评价指数的变化曲线。其中,SFFD算法的标记数据为30%。从Margin数据集下各指数的曲线变

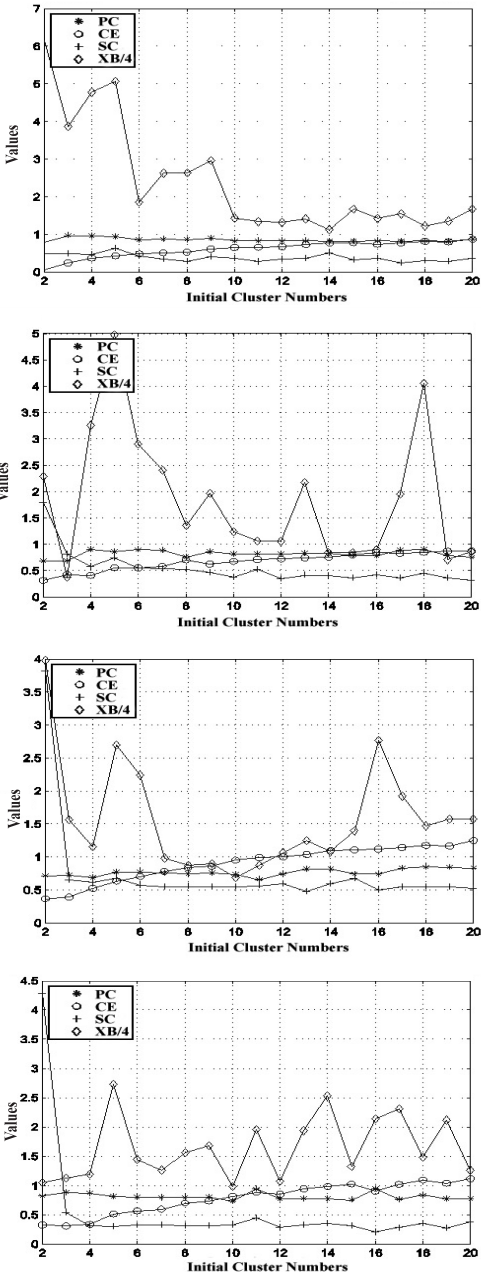


图1 不同指数下的最佳聚类数

化趋势可以看出,PC指数在  $c = 9$  时急速下跌,CE指数在  $c = 8$  时快速上升,SC指数在  $c = 11$  时处于谷底,而此时XB指数的局部最小值也是11,由于SC指数的可靠性较高,综合评估后得出最佳聚类数为11;同样的方法进行分析可知,Shape数据集下的最优聚类数为  $c = 10$ ,而Texture数据集下同样当  $c = 10$  时聚类效果最好,Combination数据集的评价结果同样是  $c = 10$ 。由于不同的特征数据集均来自于同一组树叶照片,因此,通过对4种输入数据集下的聚类结果进行模糊聚类有效性评价分析可知,该组照片的最佳聚类数为10,由于实验照片来自于10种不同的叶片图像,故该聚类评价分析结果符合研究实际。

不同特征数据集的实验结果表明:文中聚类有效

性评价算法是一种行之有效的确定聚类数的途径。

## 4 结束语

文中提出一种特征加权的半监督聚类算法,并对该算法在不同模糊聚类有效性评价算法下的聚类结果进行分析。实验结果表明,综合不同的聚类有效性评价结果,能够有效得出输入数据集的最佳聚类数,从而解决大部分聚类算法中聚类数的确定问题,具有良好的应用前景。

### 参考文献:

- [1] 许海洋,汪国安,王万森. 模糊聚类分析在数据挖掘中的应用研究[J]. 计算机工程与应用,2005,41(17):177-179.
- [2] 高新波,谢维信. 模糊聚类理论发展及应用的研究进展[J]. 科学通报,1999,44(21):2241-2251.
- [3] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社,2004:113-119.
- [4] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Intelligent Information Systems, 2001, 17(2-3):107-145.
- [5] 张惟皎,刘春煌,李芳玉. 聚类质量的评价方法[J]. 计算机工程,2005,31(20):10-12.
- [6] 李洁,高新波,焦李成. 一种基于修正划分模糊度的聚类有效性函数[J]. 系统工程与电子技术,2005,27(4):723-726.

(上接第 64 页)

繁项集结果的准确性。但由于在生成目标矩阵时,需要统计出所有的项和二次项集在事务数据库中出现的次数,从而增大了统计量,所以在时间方面没有得到优化。如何保证在减少扫描事务数据库次数的基础上提高算法的运行速度,是笔者下一步需要研究的工作。

### 参考文献:

- [1] Zaïane O R, El-Hajj M. Advances and issues in frequent pattern mining[C]//Proc of eighth Pacific-Asia conference on knowledge discovery and data mining. Sydney, Australia: [s. n.], 2004.
- [2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]//Proceedings of the ACM SIGMOD conference on management of data. Washington, D C: ACM, 1993:207-216.
- [3] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Bocca J B, Jarke M, Zaniolo C, et al. Proceeding of 20th international conference on very large data bases. Santiago, CA, USA: Morgan Kaufmann Publishers Inc, 1994:487-499.

- [7] Resson H, Wang D, Natarajan P. Adaptive double self-organizing maps for clustering gene expression profiles[J]. Neural Networks, 2003, 16(5-6):633-640.
- [8] Wu Sitao, Chow T W S. Self-organizing-map based clustering using a local clustering validity index[J]. Neural Processing Letters, 2003, 17(3):253-271.
- [9] Wu Sitao, Chow T W S. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density[J]. Pattern Recognition, 2004, 37(2):175-188.
- [10] Li Longlong, Jonathan G, He Dongjian, et al. Semi-supervised fuzzy clustering with feature discrimination[J]. Plos One, 2015, 10(9):e0131160.
- [11] 李龙龙,王美丽. 基于加权二叉树的自适应遗传算法研究[J]. 计算机技术与发展, 2010, 20(11):95-99.
- [12] 李洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报, 2006, 34(1):89-92.
- [13] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. [s. l.]:Springer, 1983.
- [14] Bensaid A M, Hall L O, Bezdek J C, et al. Validity-guided (re)clustering with applications to image segmentation[J]. IEEE Transactions on Fuzzy Systems, 1996, 4(2):112-123.
- [15] Xie X L L, Beni G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8):841-847.

- [4] 程玉胜,邓小光,江效尧. Apriori 算法中频繁项集挖掘实现研究[J]. 计算机技术与发展, 2006, 16(3):58-60.
- [5] 郭福亮,左凯伶. 关联规则挖掘中 Apriori 算法的一种改进[J]. 计算机与数字工程, 2007, 35(5):3-4.
- [6] Han Jiawei, Kamber M. Data mining: concepts and techniques[M]. 3th ed. Beijing: China Machine Press, 2012.
- [7] 李小兵,吴锦林,薛永生,等. 关联规则挖掘算法的改进与优化研究[J]. 厦门大学学报:自然科学版, 2005, 44(4):468-471.
- [8] 包震宇. 基于粗糙集对 Apriori 算法的改进[D]. 上海:上海师范大学, 2010.
- [9] 马晓辉. 一种基于关联规则 Apriori 算法的改进研究[J]. 现代计算机, 2011(6):6-8.
- [10] 吕桃霞,刘培玉. 一种基于矩阵的强关联规则生成算法[J]. 计算机应用研究, 2011, 28(4):1301-1303.
- [11] 张笑达,徐立臻. 一种改进的基于矩阵的频繁项集挖掘算法[J]. 计算机技术与发展, 2010, 20(4):93-96.
- [12] 芦洁,刘志健. 挖掘关联规则中对 Apriori 算法的一个改进[J]. 微电子学与计算机, 2006, 23(2):10-12.
- [13] 顾琳,黎敬涛,张兴涛. 对 Apriori 算法的一种改进—基于 0-1 矩阵处理算法[J]. 电脑知识与技术, 2007, 4(21):814-816.